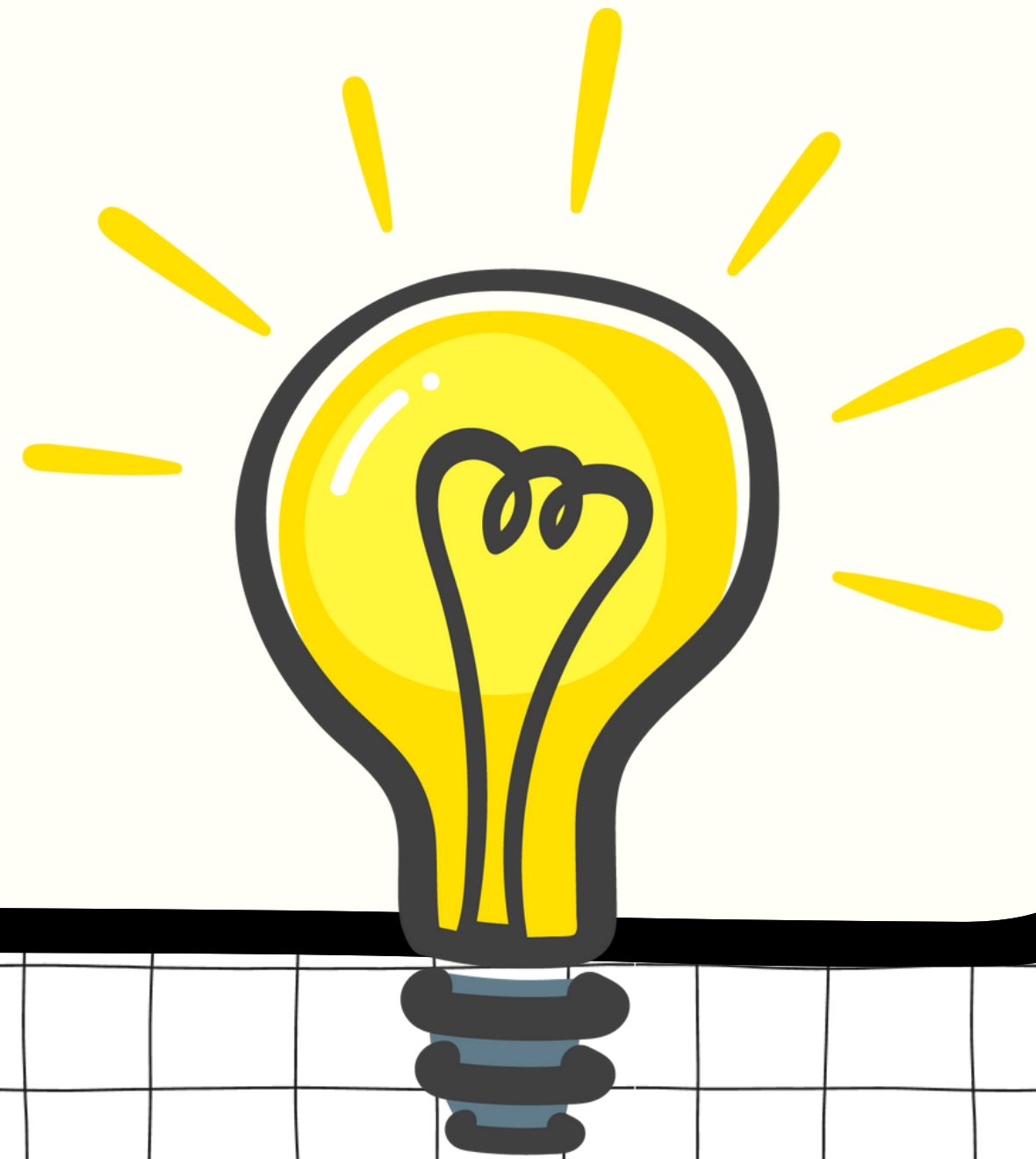
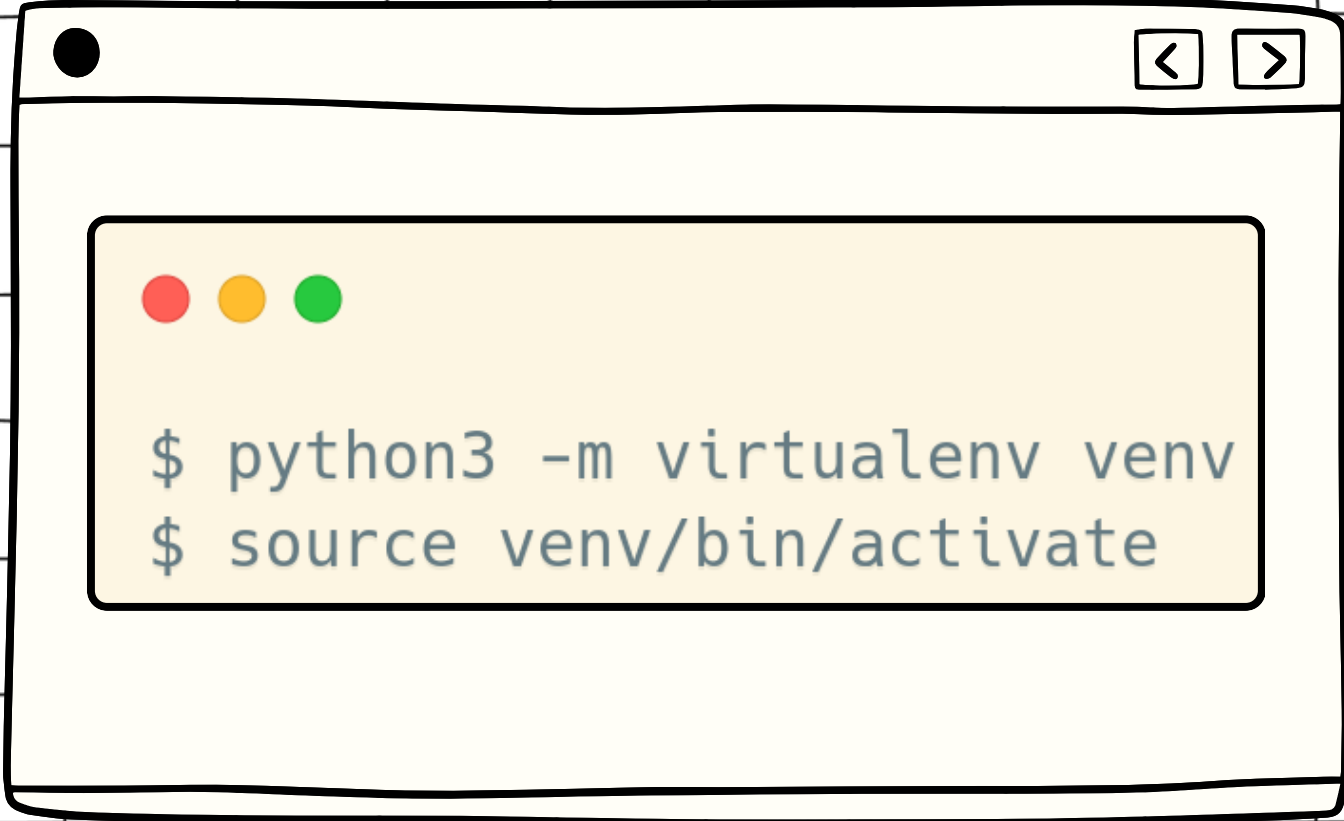


Scraping Data OLX

Furqan Al Ghifari Zulva (2108107010053)



1. Membuat Virtual Environment



A hand-drawn illustration of a terminal window. It has a title bar with a close button (a circle) and two navigation buttons (left and right arrows). Inside the window is a light yellow rectangular area representing the terminal output, which contains three colored dots (red, yellow, green) and two lines of code. The background of the slide features a grid pattern with yellow decorative shapes in the corners.

```
$ python3 -m virtualenv venv  
$ source venv/bin/activate
```



A hand-drawn illustration of a terminal window, similar to the one above. It has a title bar with a close button (a circle) and two navigation buttons (left and right arrows). Inside the window is a light yellow rectangular area representing the terminal output, which contains three colored dots (red, yellow, green) and one line of code. The background of the slide features a grid pattern with yellow decorative shapes in the corners.

```
$ pip install scrapy
```

2. Mengkonfigurasi Project Scrapy



```
# scrapy startproject <project_name>  
scrapy startproject bigdatascraper
```



```
# syntax is --> scrapy genspider <name_of_spider> <website>  
scrapy genspider bigdataspider https://www.olx.co.id/motor-bekas_c200
```



3. Menemukan CSS Selector dengan Menggunakan Scrapy Shell

```
$ scrapy shell
```

```
[s] Available Scrapy objects:
[s] scrapy      scrapy module (contains scrapy.Request, scrapy.Selector, etc)
[s] crawler     <scrapy.crawler.Crawler object at 0x7fd94b512980>
[s] item        {}
[s] settings     <scrapy.settings.Settings object at 0x7fd94b5131f0>
[s] Useful shortcuts:
[s] fetch(url[, redirect=True]) Fetch URL and update local objects (by default, redirects are followed)
[s] fetch(req)          Fetch a scrapy.Request and update local objects
[s] shelp()             Shell help (print this help)
[s] view(response)      View response in a browser
2023-11-07 04:42:06 [asyncio] DEBUG: Using selector: EpollSelector
In [1]:
```

mengambil halaman website

```
In [1]: fetch('https://www.olx.co.id/motor-bekas_c200')
2023-11-07 04:46:50 [scrapy.core.engine] INFO: Spider opened
2023-11-07 04:46:51 [scrapy.core.engine] DEBUG: Crawled (200)
```

```
In [2]: response
Out[2]: <200 https://www.olx.co.id/motor-bekas_c200>
```

```
In [3]: response.xpath('//li/a/div[contains(@class, "fTzt3")]')
```

```
Out[3]:
[<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><span class="_2Ks6...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>',
<Selector query="//li/a/div[contains(@class, "fTzt3")]> data='<div class="fTzt3"><div class="_1IpS4...>']
```


mengambil halaman website

The screenshot shows a web browser displaying the OLX website, specifically the 'motor-bekas_c200' page. The browser's address bar shows the URL `olx.co.id/motor-bekas_c200`. The page features a sidebar with categories like 'Motor', 'Motor Bekas (155.322)', 'Aksesoris (7.411)', and 'Helm (2.983)'. The main content area displays three motorcycle listings, each with a 'HIGHLIGHT' tag, an image, a price, and a year. The first listing is a 'Yamaha R15 v3 Upgrade Nego' for Rp 25.000.000, 2017. The second is a 'Yamaha Mio M3 125 (BARU)' for Rp 17.500.000, 2023. The third is a 'Volta 401 Motor Listrik (Bekas)' for Rp 5.000.000, 2022. A Chrome DevTools overlay is visible at the bottom, showing the 'Elements' panel with the HTML structure of the first listing. The HTML structure is as follows:

```
<div class="...">
  <div class="..._1xlea">
    <div>
      <ul class="..._10aCo" data-aut-id="itemsList">
        <li data-aut-id="itemBox" data-aut-category-id="200" class="..._1DNjI">
          <a class="..._2cbZ2" href="/item/jual-cepat-yamaha-r15-v3-upgrade-nego-iid-908200030">
            <figure class="..._3Urc5" data-aut-id="itemImage">
              <div class="ftZT3">
                <div class="..._1IpS4">
                  <span class="..._2Ks63" data-aut-id="itemPrice">Rp 25.000.000</span>
                  <span class="..._YBbhy" data-aut-id="itemDetails">2017</span>
                  <span class="..._2poNJ" data-aut-id="itemTitle">JUAL CEPAT Yamaha R15 v3 Upgrade Nego</span>
                </div>
              </div>
            </a>
          </li>
        </ul>
      </div>
    </div>
  </div>
```

mendapatkan konten

```
In [7]: products = response.xpath('//div[contains(@class, "fTzT3")]')
```

```
In [8]: len(products)
```

```
Out[8]: 20
```

```
In [9]: products
```

```
Out[9]:
```

```
[<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><span class="_2Ks6...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>',  
<Selector query='//div[contains(@class, "fTzT3")]' data='<div class="fTzT3"><div class="_1IpS4...>']
```

...
mengeksrak **harga**, **tahun**,
brand, dan **alamat** untuk
setiap **product** didalam
list **products**

```
In [10]: product = products[0]

In [11]: product
Out[11]: <Selector query='//div[contains(@class, "fTZT3")]' data='<div class="fTZT3"><div class="_1IpS4...>'

In [12]: product.css('span._2Ks63::text').get()
Out[12]: 'Rp 160.000.000'

In [13]: product.css('span._2Ks63::text').extract()
Out[13]: ['Rp 160.000.000']

In [14]: type(product.css('span._2Ks63::text').get())
Out[14]: str

In [15]: type(product.css('span._2Ks63::text').extract())
Out[15]: list

In [16]: len(product.css('span._2Ks63::text').extract())
Out[16]: 1
```

```
In [18]: product.css('span._2Ks63::text').get()
Out[18]: 'Rp 160.000.000'

In [19]: product.css('span.YBbhy::text').get()
Out[19]: '2023'

In [20]: product.css('span._2poNJ::text').get()
Out[20]: 'Honda Monkey with Sidecar'

In [21]: product.css('span._2VQu4::text').get()
Out[21]: 'Pesanggrahan, Jakarta Selatan'
```



```
In [26]: product2 = products[1]
```

```
In [27]: product2
```

```
Out[27]: <Selector query='//div[contains(@class, "fTZT3")]' data='<div class="fTZT3"><div class="_1IpS4...>
```

```
In [28]: product2.css('span._2Ks63::text').get()
```

```
Out[28]: 'Rp 30.800.000'
```

```
In [29]: product2.css('span.YBbhy::text').get()
```

```
Out[29]: '2018'
```

```
In [30]: product2.css('span._2poNJ::text').get()
```

```
Out[30]: 'Vespa s 2018 iget km br 11 gres rawatan TT matic taun muda gas'
```

```
In [31]: product2.css('span._2VQu4::text').get()
```

```
Out[31]: 'Pesanggrahan, Jakarta Selatan'
```

```
In [32]: product3 = products[2]
```

```
In [33]: product3
```

```
Out[33]: <Selector query='//div[contains(@class, "fTZT3")]' data='<div class="fTZT3"><div class="_1IpS4...>
```

```
In [34]: product3.css('span._2Ks63::text').get()
```

```
Out[34]: 'Rp 14.500.000'
```

```
In [35]: product3.css('span.YBbhy::text').get()
```

```
Out[35]: '2023'
```

```
In [36]: product3.css('span._2poNJ::text').get()
```

```
Out[36]: 'Yamaha New Mio Gear 2023 standar Merah'
```

```
In [37]: product3.css('span._2VQu4::text').get()
```

```
Out[37]: 'Pancoran, Jakarta Selatan'
```

4. Mengupdate Kode Spider

```
bigdataspider.py X olxproduct.json
bigdatascraper > bigdatascraper > spiders > bigdataspider.py > BigdataspiderSpider > parse
1  import scrapy
2
3
4  class BigdataspiderSpider(scrapy.Spider):
5      name = "bigdataspider"
6      allowed_domains = ["www.olx.co.id"]
7      start_urls = ["https://www.olx.co.id/motor-bekas_c200"]
8
9      def parse(self, response):
10         products = response.xpath('//li/a/div[contains(@class, "fTZT3")]')
11         for product in products:
12             yield{
13                 'brand' : product.css('span._2poNJ::text').get(),
14                 'lokasi' : product.css('span._2VQu4::text').get(),
15                 'tahun' : product.css('span._YBbhy::text').get(),
16                 'harga' : product.css('span._2Ks63::text').get()
17             }
18
```

5. Menjalankan Spider

```
$ scrapy crawl bigdataspider -o olxproduct.json
```

```
bigdataspider.py  olxproduct.json x
bigdatascraper > {} olxproduct.json > {} 18 > lokasi
1  [
2  {"brand": "Yamaha Xmax 2022 Hitam Low KM Full Orisinal", "lokasi": "Pesanggrahan, Jakarta Selatan", "tahun": "2022", "harga": "Rp 55.500.000"},
3  {"brand": "YAMAHA NEW NMAX ABS CONNECTED 2021", "lokasi": "Pancoran, Jakarta Selatan", "tahun": "2021", "harga": "Rp 26.500.000"},
4  {"brand": "Moge Kawasaki ZX6R ZX636 Anniversary Series", "lokasi": "Cilandak, Jakarta Selatan", "tahun": "2014", "harga": "Rp 257.000.000"},
5  {"brand": "YAMAHA MIO MATIC WARNA HITAM TAHUN 2010 KONDISI MULUS PAJAK HIDUP.", "lokasi": "Pancoran, Jakarta Selatan", "tahun": "2010", "harga": "Rp 5.100.000"},
6  {"brand": "(DP 1 Jt) Beat CBS ISS Hitam 2022 Cash & Kredit", "lokasi": "Kebayoran Lama, Jakarta Selatan", "tahun": "2022", "harga": "Rp 14.900.000"},
7  {"brand": "Honda Supra X Helm In 2011", "lokasi": "Kebayoran Lama, Jakarta Selatan", "tahun": "2011", "harga": "Rp 5.250.000"},
8  {"brand": "REVO CW FI TH 2017'ANGSURAN 300rbn KTP DAERAH BSA CASH /TT/CC BUNGA 0%", "lokasi": "Pesanggrahan, Jakarta Selatan", "tahun": "2017", "harga": "Rp 8.500.000"},
9  {"brand": "REVO CW FI TH 2017[ANGSURAN 300rbn KTP DAERAH BSA CASH TT CC 0%SPAY", "lokasi": "Pasar Minggu, Jakarta Selatan", "tahun": "2017", "harga": "Rp 8.500.000"},
10 {"brand": "REVO CW FI TH 2017;ANGSURAN SUPER RINGAN 300rbn KTP DAERAH BSA CASH/TT", "lokasi": "Mampang Prapatan, Jakarta Selatan", "tahun": "2017", "harga": "Rp 8.500.000"},
11 {"brand": "REVO CW FI 2017*ANGSURAN 300rbn KTP DAERAH BSA CASH/ TT CC BUNGA 0%", "lokasi": "Kebayoran Lama, Jakarta Selatan", "tahun": "2017", "harga": "Rp 8.500.000"},
12 {"brand": "REVO CW FI 2017:ANGSURAN 300rbn KTP DAERAH BSA CASH/TT CC 0%/SPAY", "lokasi": "Cilandak, Jakarta Selatan", "tahun": "2017", "harga": "Rp 8.500.000"},
13 {"brand": "Yamaha Mio Soul Gt115 Injeksi 2013 Pancoran", "lokasi": "Pancoran, Jakarta Selatan", "tahun": "2013", "harga": "Rp 6.200.000"},
14 {"brand": "Yamaha Mio Soul Gt115 Injeksi 2013 Mampangprapatan", "lokasi": "Mampang Prapatan, Jakarta Selatan", "tahun": "2013", "harga": "Rp 6.200.000"},
15 {"brand": "Yamaha Mio Soul Gt115 Injeksi 2013 Kebayoranlama", "lokasi": "Kebayoran Lama, Jakarta Selatan", "tahun": "2013", "harga": "Rp 6.200.000"},
16 {"brand": "For Sale Honda New Cbr 150 Facelift Cash/Kredit", "lokasi": "Mampang Prapatan, Jakarta Selatan", "tahun": "2018", "harga": "Rp 17.800.000"},
17 {"brand": "Vespa Primavera 50th Anniversary 2018 Limited Edition - Low KM", "lokasi": "Kebayoran Baru, Jakarta Selatan", "tahun": "2018", "harga": "Rp 45.900.000"},
18 {"brand": "Honda New ADV 160 CBS Very Low Kilometer", "lokasi": "Pasar Minggu, Jakarta Selatan", "tahun": "2022", "harga": "Rp 33.500.000"},
19 {"brand": "Motor Custom Basic Byson", "lokasi": "Kebayoran Baru, Jakarta Selatan", "tahun": "2012", "harga": "Rp 17.000.000"},
20 {"brand": "DIJUAL VESPA EXCEL 150 TAHUN 1991", "lokasi": "Pasar Minggu, Jakarta Selatan", "tahun": "1991", "harga": "Rp 25.500.000"},
21 {"brand": "Peugeot scooters django ABS", "lokasi": "Mampang Prapatan, Jakarta Selatan", "tahun": "2022", "harga": "Rp 65.000.000"}
22 ]
```



감사합니다
Thank you